Class Imbalance Problems with 4-D interpolator.

Outline:

1. How we discovered the problem.
2. Define the Problem
3. Define some terms used in Literature
4. Simulated example of Class Imbalance effect using linear regression
5. Overview of Class Imbalance solutions suggested in literature
6. Review of our Plans to address problem.

Identifying Class Imbalance

Class Imbalance is a term for imbalance between data used for training and data used for prediction.

It is identified by comparing the distribution of the independent variables in the training data to the distribution of independent variables in the prediction space or target data.

If a region of the training data has a much higher or much lower density of observations than the target data, then Class Imbalance exists.

For the 4D-interpolator

Training data =

Fixed Station Data,

Citizen Monitoring

High Frequency Data (ConMon, DataFlow, Vertical Array)

Prediction Space or Target Space = The interpolation grid

Waterbody Length (1 km)

Waterbody Breadth (1 km)

Sample Depth (1 m)

Time (primary 1 day, secondary 1 hour)

Region with Training data > Target Data defines Majority Class

Region with Training data < Target Data defines Minority Class

When Class Imbalance becomes a problem.

Least Squares training adjusts model coefficients to Minimize Prediction Error over all Training data.

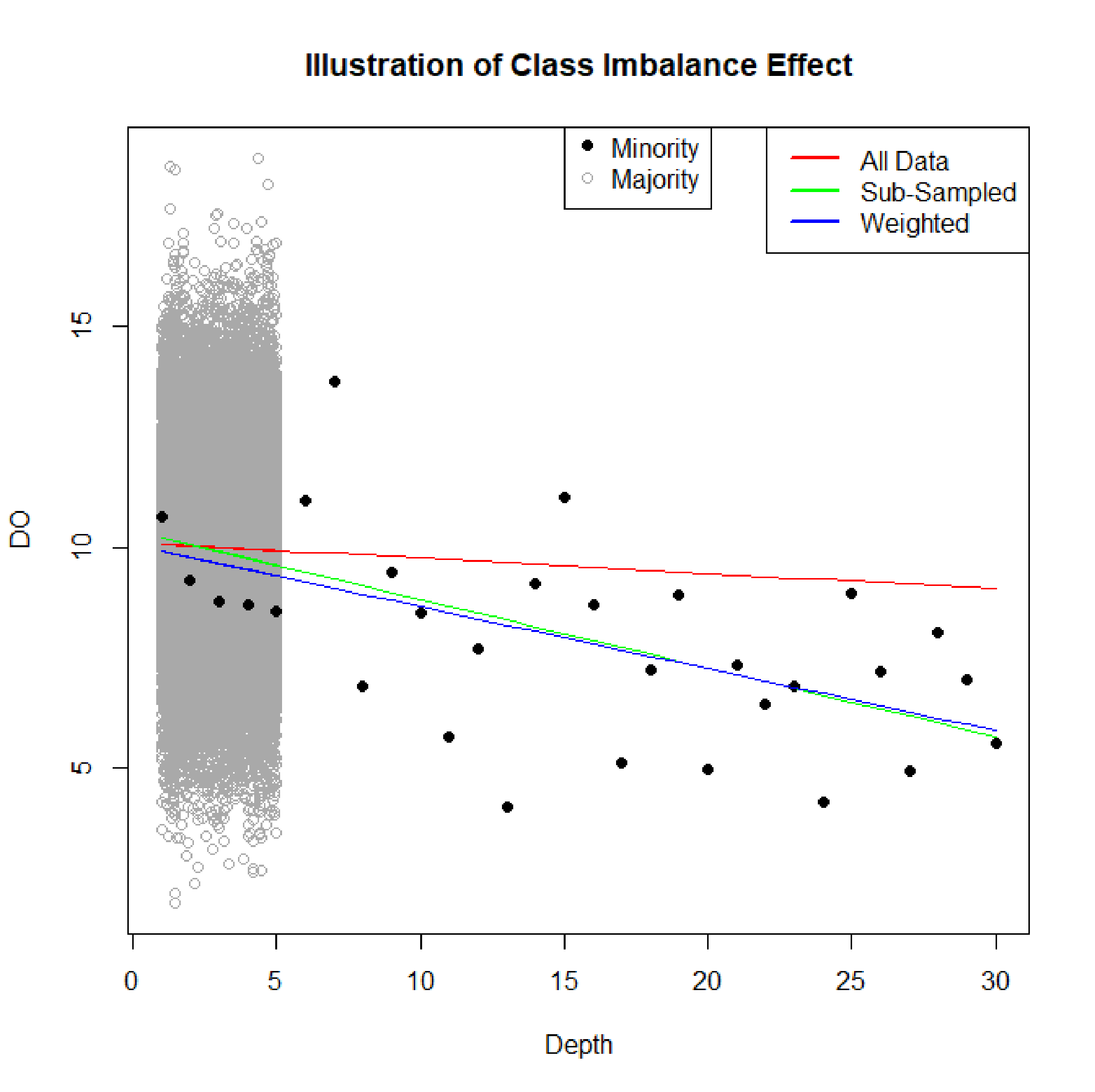
(small improvement) * (majority class) > (large degradation) * (minority class)

However, Minority Class represents a large region in the prediction space.

Can lead to poor prediction in a large part of the prediction space.

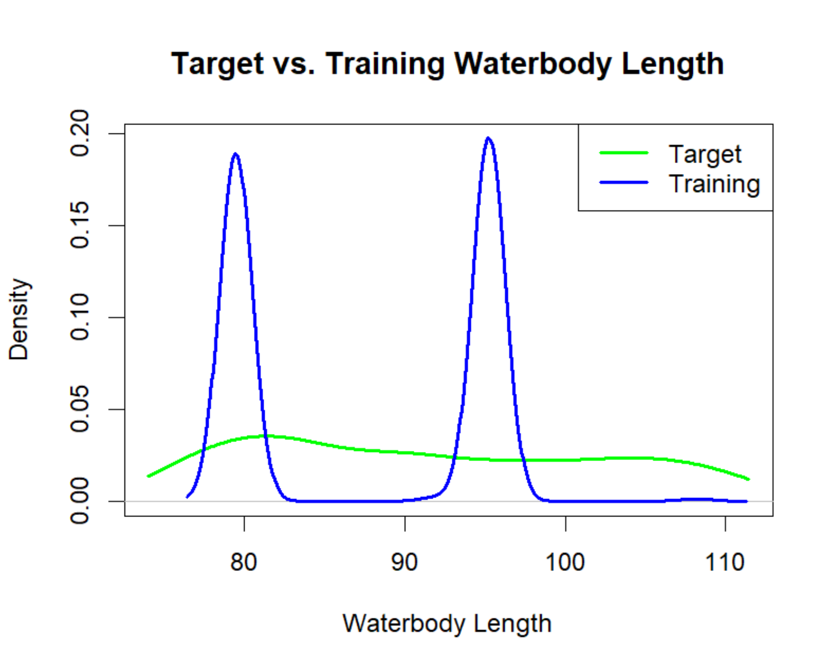
Class Imbalance becomes a problem because model training methods (e.g. least squares) optimize the model fitting criterion, (e.g. minimizing the sum of squared errors) over the entire data set. The model training procedure will naturally tend to minimize error in the region with abundant data at the expense of poorer prediction in regions with less data. This impairment of model performance in regions with less data becomes critical when regions with less data require better prediction for accurate assessment.

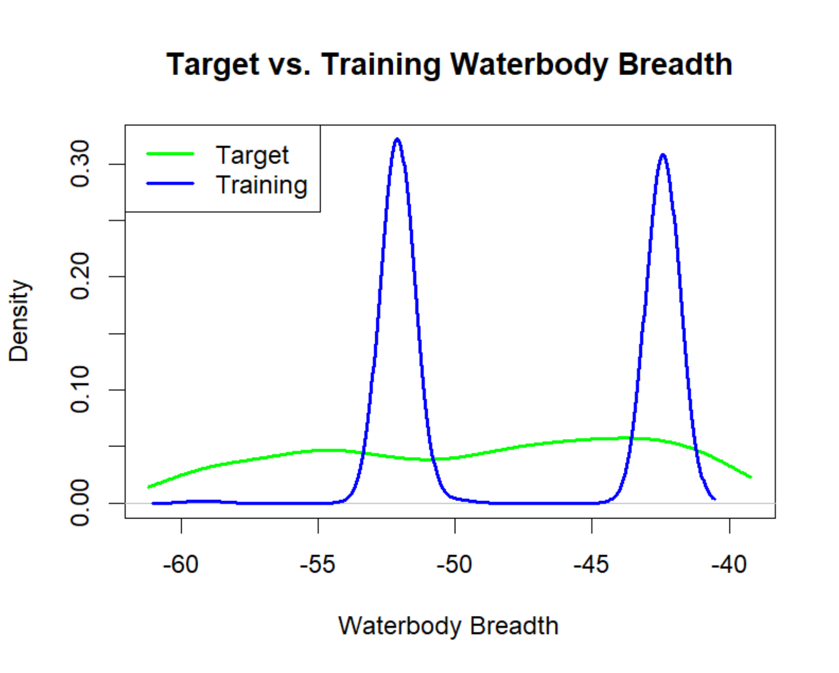
Rebecca’s examples for YRKMH

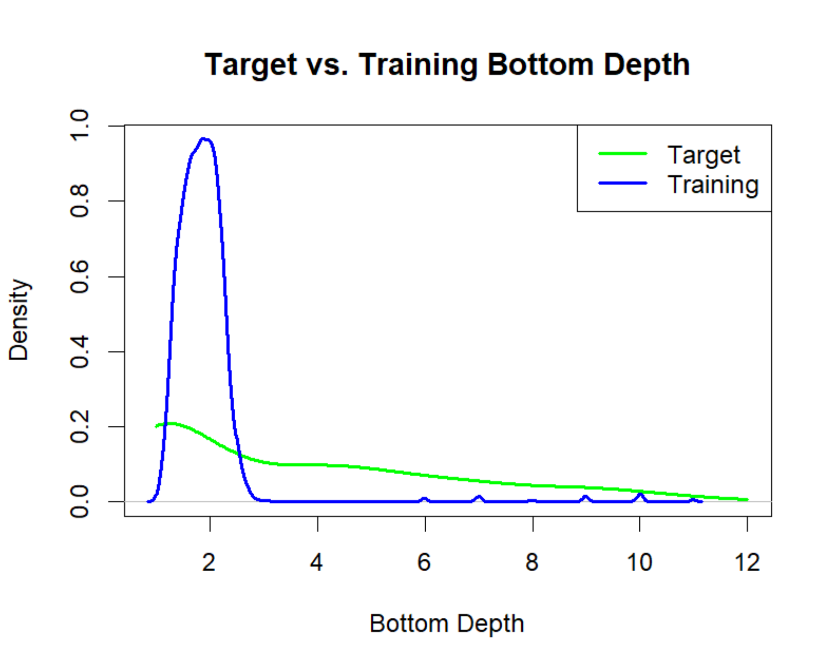


	Sample Size		Weights		RMSE	
	Majority	Minority	Majority	Minority	Majority	Minority
Model 1 (red)	40001	30	1	1	2.0055	2.0755
Model 2 (green)	200	30	1	1	2.0001	1.9009
Model 3 (blue)	200	30	5 / 205	1	2.0002	1.8913

The addition of ConMon and DataFlow creates class imbalance in the 4D training data







(Training > Target) defines Majority Class (Training < Target) defines Minority Class

2000s Data Science and Classification Models, Class Imbalance

Class Imbalance is often defined by the end point of interest.

Examples:

- **Mortgage default (majority class – non-default, defaults of interest)**
- **Rare disease (majority class = healthy population, disease of interest)**
- **Fraudulent Credit Card Transaction (majority = legitimate transactions, fraud of interest)**
- **Detecting Bots vs Real People on Websites (majority = real people, bots of interest)**

Tools for restoring Balance from literature

Useful Corrective Measures:

Experimental units – eliminate Pseudo-replication

Undersampling, Reduce the number of observations in the majority class.

Class Weights, Assign higher weights to the minority class during training to balance the influence of each class.

Not Useful Corrective Measures:

Oversampling, Sampling with replacement to increase the number of instances in the minority

Threshold Adjustment, Modify the decision threshold for classifying instances to improve sensitivity for the minority class.

Sizes of Experimental Units

Primary Units are defined by interpolation grid –

Waterbody Longitude (1km)

Waterbody Latitude (1km)

Sample Depth (1m)

Time (1 day)

Modeled by Mid-Day Means Model + correlated errors

Secondary Units nested within Primary Units –

Time (1 hour)

Modeled by small scale model of diel and tidal cycles + correlated errors

Under-sampling or sub-sampling

- ConMon and Vertical Array
  - 1 obs per hour
- DataFlow
  - 1 obs per 500m

Down weighting majority class data for Daily Means model.

.

- ConMon and Vertical Array
  - Wgt = 1/24
- DataFlow
  - Wgt = ??

Questions?

Historical Notes: Move to end

1970s, Sampling Theory, Unequal Probability Sampling – Cochran

Lead so weighted estimation.

1980 Experimental Design, Pseudo-replication – Hurlburt, 1980

Leads to averaging or subsampling.

1980s Milliken and Johnson

Sizes of Experimental Units, Nested Designs, Hierarchical Designs

Keep from here down.

2000s Data Science and Classification Models, Class Imbalance

Class Imbalance not just based on Probability, but can be based on Utility.

Examples:

- Mortgage default (majority class – non-default)
- Rare disease (majority class = healthy population)
- Fraudulent Credit Card Transaction (majority = legitimate transactions)
- Detecting Bots vs Real People on Websites (majority = real people)